

## A stopping criterion for Markov chains when generating independent random graphs

J. RAY\*, A. PINAR, C. SESHADHRI  
Sandia National Laboratories, Livermore, CA, USA  
Email: {jray, scomand, apinar}@sandia.gov

[Received on 21 July 2014]

Markov chains are convenient means of generating realizations of networks with a given (joint or otherwise) degree distribution, since they simply require a procedure for rewiring edges. The major challenge is to find the right number of steps to run such a chain, so that we generate truly independent samples. Theoretical bounds for mixing times of these Markov chains are too large to be practically useful. Practitioners have no useful guide for choosing the length, and tend to pick numbers fairly arbitrarily. We give a principled mathematical argument showing that it suffices for the length to be proportional to the number of desired number of edges. We also prescribe a method for choosing this proportionality constant. We run a series of experiments showing that the distributions of common graph properties converge in this time, providing empirical evidence for our claims.

*Keywords:* graph generation; Markov chain Monte Carlo; independent samples

### 1. Introduction

Graphs are a common topological representation across a variety of scientific fields. They are used when relations between a large number of entities have to be specified in a succinct manner. Chemical reactions, molecules, social networks (both physical and online) and the electric grid are some common examples. In many cases, we may have partial information about the graph, requiring generation of many graphical realizations consistent with the (partial) characterization. Such situations arise in case of large online social networks (of which only a small part can be sampled tractably) or where the data simply cannot be collected e.g., the web of human sexual relations (which only allows the estimation of the degree distribution). In other cases, privacy concerns can prevent the distribution of a graph for experimentation and or study, e.g., networks of email communications, critical infrastructure nets, etc. This gives rise to the problem of constructing “sanitized” proxies, that preserve only some properties of the original graph. Thus, being able to generate independent graphs conditioned on an incomplete set of graphical measurements is essential for many applications.

Many graph generation methods aim to preserve the salient features of graphs [1, 7, 9, 11, 12, 15, 19, 24, 29]. For statistical analysis, we need algorithms that can generate uniformly random instances from the space of graphs with a specified feature. There has been significant work on random graph generation with a given degree distribution (DD), which specifies the number of vertices with a given degree. In [14], the problem of generating a graph with a given degree distribution was reduced to a perfect matching problem, which can be used to generate random instances by employing results in sampling perfect matchings [6, 13]. Alternatively, sequential sampling methods were investigated

\*Corresponding author

in [3, 5]. These methods can be compute intensive, and in practice, Markov chains (MC) are widely used due to their simplicity and flexibility. The MC is started using a graph that honors the specified graphical characteristics, and uses a procedure that can “rewire” a graph, to create ensembles of graphs with the same degree distribution. Taylor [28] showed that edge-swaps could modify a graph while preserving its DD. Kannan et al [14] analyzed the mixing time of such a MC, whereas Gkantsidis [10] devised a MC scheme that avoids self loops. In [27], Stanton and Pinar used an MC to generate an ensemble of graphs using a rewiring scheme that preserved the joint degree distribution (JDD), which specifies the number of edges between vertices of specified degrees. Stanton and Pinar also empirically analyzed the mixing of the MC using the binary “time-series” traced out by the appearance/disappearance of edges between two labeled nodes, as the MC executed its random walk in the space of graphs. They showed that the autocorrelation of the time-series decayed to zero, a necessary condition for the MC to converge to its stationary distribution [25].

Graph rewiring schemes that preserve DD or JDD are simple (and fast) and MC chains driving them are easy to construct. However, successive graphs generated by the MC are only slightly different and the MC has to proceed for a large number of steps  $N$  before the initial graph is “forgotten.” The mixing time estimates in [14] take the form of upper bounds and of  $O(|E|^6)$ , where  $|E|$  is the number of edges in a graph. Even for small graphs with 1000 edges, the bound is intractable. In practice,  $N$  is chosen arbitrarily. Failure to mix completely leads to the generation of correlated samples and any results derived from them are erroneous. While the method in [27] demonstrated the use of autocorrelation to establish mixing, it is an *a posteriori* test which provides no *a priori* guidance regarding the length of the MC chain. Further, the method requires a small, user-specified threshold, below which the time-series autocorrelation is deemed zero.

In this paper, we construct analytical models that estimate  $N$ . These models track the evolution of the binary “time-series” formed by the edges of labeled graphs generated by the MC. We test the model under conditions where the DD or the JDD is held constant. In Sec. 2, the model is used to derive an expression for the mixing time based on when the binary time-series begins to resemble independent draws from a distribution. The models predict that the mixing time is proportional to  $|E|$ , the number of edges in the graph. The model holds true for a “representative” edge and exploits the constancy of DD (or JDD) in arriving at its prediction. The model is tested with real graphs in Sec. 3. In Sec. 4, we develop a data-driven method, that assumes neither a constant DD (or JDD) nor a representative edge, to investigate the independence of the edge time-series. We use this test to verify the assumptions made when developing mixing time expressions in Sec. 2. We determine the fraction of the edges for which the model prediction of  $N$  is an underestimate, and the consequences of the lack of stationarity on the ensemble of graphs sampled in this manner.

## 2. Theoretical analysis

The goal of this paper is to provide a mathematically principled argument for running a MC  $O(|E|)$  steps to generate independent graphs with  $|E|$  edges. The constant hidden in the big-Oh depends on the desired accuracy. The graphs so generated may have a prescribed DD or a JDD; we find that a MC on graphs with a prescribed DD mixes slightly easier than those where the JDD is preserved. Our empirical results show that  $5|E| - 30|E|$  steps are sufficient for mixing these MCs. We provide a mathematical justification for this observation.

Consider a MC on the space of graphs and two labeled vertices  $u$  and  $v$ . Under certain circumstances, the existence of an edge  $(u, v)$  can also be described by a Markov chain. Thus we model the behavior of the MC on the space of graphs in terms of a set of coupled 2-state Markov chains, each representing